Report on the current state of the research for the Ph.D. thesis and request for extension of submission deadline

David García Garzón, PhD student in TICMA program.

1 Introduction

This document reports the status of the work of the PhD candidate in order to ask for an extension to the submission of the thesis up to the end of the second trimester of the 2011/2012 course.

2 Current status

This research has been performed during years 2008 - 2011. The proposal was entitled "Relating audio and 3D scenarios in audiovisual productions" which was presented and approved in June 2008.

The focus of that research are technologies that automate the production of audiovisual productions containing 3D audio, including audio rendering, physical coding, 3D decoding and environmental acoustic inference.

The work done during this last year covers the areas of analysing the 3D decoding and room inference.

2.1 3D audio decoding

The first block of topics are related to 3D audio decoding. There is already a lot of previous literature in this topic so we have centered on several concrete aspects that has not received enough attention in existing literature. One is normalizing and compensating HRTF database for their use for simulating 3D ambisonics decodings in binaural setups. The other, and this is the bigger topic of our research, is analyzing the filtering effect that ambisonics introduces for higher frequencies even for high order decodings.

2.1.1 Quality analysis of Ambisonics decoding

In Ambisonics decoding, all loudspeakers participate of reproduction even for highly directional soundfields. The presence of the listener's head at the sweetspot introduces different shadows and delays to the signals arriving from each loudspeaker, thus disturbing the acoustic field reconstruction at both ears. Although this does not necessarily affect localization cues, it does create a filter which is clearly noticeable, specially when reproducing

frequency-rich content like music. Such degradation is also present in binaural signals obtained from Ambisonics, a procedure which simulates a perfect setup in terms of loudspeakers and listener placement.

We have quantified this effect and its dependence on parameters such as the Ambisonics order and decoding scheme used, and the listener's head radius. Exact analytic expressions for the filters generated in 2D and 3D speaker arrays have been obtained, for frequency and time domains, under some approximations that still capture the main features of the empirically measured filters. We still need to empirically evaluate the predictions from the analytical model that we didn't measure.

A paper covering this research is almost ready for submission, and, it will be the main body the thesis, so we would like to incorporate some of the early feedback we will receive from the reviewers. The current working draft of the paper, is attached to this repport as annex. It has the following structure:

1. **Formalism of Ambisonics to binaural conversion:** We introduce mathematical formalisms required for the rest of the paper.
   (a) **Ambisonics decoding**: We obtain general expressions to decode ambisonics using 2D or 3D regular arrays of speakers of any order.
   (b) **Ambisonics to binaural**: By formalizing the standard conversion from ambisonics to binaural, we get to the concept of HRTF field and its spherical harmonic expansion which are quantities that appear naturally when doing such conversions.
2. **Analytical expressions for the HRTF spherical harmonics**: By providing a given set of simplifications, we get analytical expressions for the spherical hamonic components of the HRTF field in time and frequency domains for 3D and 2D spherical harmonics components.
3. **Perceived filter for plane wave decodings:** By using the previous components, we obtain the perceived filters when decoding a plane wave. Such filter depends on the incoming orientation and is quite different depending on the decoding criteria.
4. **Realistic conditions:** We release the simplifications we did to obtain the analytical expressions and see whether the conclusions still stand.

The results of this research state that 3D Ambisonics sensitively deteriorates the high frequencies, even for high order Ambisonics, but the analytical results point to some feasible compensations strategies. We would like to develop and evaluate such strategies by means of analytical evaluation and user tests. Equalization and evaluation might lead to a new paper to be written.

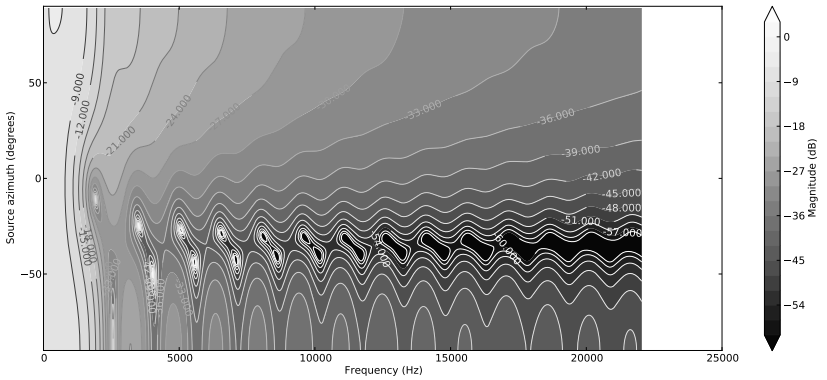

Figure 1: Filter variation depending on the orientation of the plane wave relative to the head of a 1st order $\max_{rE}$ 3D decoding. -90 degrees means ipsilateral to the considered ear, 90 is contralateral and 0 degrees, front or back.

2.1.2 Gathering, standardization and compensation of existing HRTF databases

Head Related Transfer Functions (HRTF) capture the impulse response of a sound at the ear entrance coming from a point at a given direction related to the head. HRTF's are used for binaural and transaural 3D audio exhibition as well as for simulating actual hearing of other exhibition systems. HRTF can be obtained by recording, by simulation or, in very simplified cases, by analytical formulas.

In any case, HRTF performance as to localization heavily depends on how much the original head they were obtained from matches the listener's one. Many HRTF databases are available for research but they very heterogeneous regarding storage format, quality, postprocessing, coordinate system conventions, angular sampling and others.

A first work of this research was to gather such databases and make them available under a common set of conventions and formats, so that subsequent experiments can easily switch from one database to another.

Several databases were collected including the ones from [1], [2] and [3]. They were recoded in the same way regarding format (wav), sampling rate, and indexes were generated so that the collect the mapping between each audio file to its orientation.

Also a procedure was defined so that given an HRTF database its decomposition in spherical harmonics can be obtained provided several strategies to compensates the irregularities that deviate them from a continuous infinitesimal array of speakers. The strategies included are:

- Vbap generation of missing hrtfs for the lower cap, which is often missing
- Ponderation by the solid angle which includes orientations which are nearer to each real speaker

2.2 Scene inference from audio (SIFA)

The *scene inference from audio* (SIFA) problem consists on obtaining a set of room parameters regarding its geometry and materials so that its acoustics properties are undistinguishable from the ones captured in a room impulse response or in a echoic recording.

This is in contrast to the common practice of generating an impulse response from either a real-world space, or from a virtual space. In other words: rather than generating an acoustic impulse response from a given room, we generate a room from a given impulse response (IR); and hence this task is called the "Inverse IR" problem.

Indeed, the research performed attempts at inferring not only the room geometry and the property of the materials present in it, but also the position of the sound sources and receivers. We will therefore refer to such as Scene Inference From Audio (SIFA) algorithms, whereby the concept of scene refers to the more generic setup described above. Such strategies are implemented but pending of user tests.

2.2.1 Problem formalization

To formalize the problem we define a space of scenes $S = \{\mathbf{G2}, \mathbf{c}, \mathbf{s}, \mathbf{r}\}$, as the space of all possible closed two-dimensional surfaces $G2$, of all possible frequency-dependent coefficients c that characterize the acoustics of points within the surfaces, and all possible locations of a sound source $\mathbf{s} = (s_x, s_y, s_z)$, and a receiver, $\mathbf{r} = (r_x, r_y, r_z)$, in the interior of such surfaces.

We have greatly simplified the problem and analyze its main features. First, the infinite-dimensional space G2 has been restricted to the space of all possible rectangular rooms, whereby all walls are orthogonal to the three Cartesian axes; such rooms are fully characterized by specifying its three dimensions $\mathbf{l} = (l_x, l_y, l_z)$. Second, the space of coefficients c has been restricted to one real number, $R \in [0, 1]$, describing the reflection coefficient, being the same for all walls. With these simplification, the space of scenes S can be parametrized by the 10 real numbers $\{R, \mathbf{l}, \mathbf{s}, \mathbf{r}\}$.

The parametrization of $\mathbf{S}$ can still be optimized by making use of symmetries to reduce redundancy. In particular note that the following three operations lead to acoustically indistinguishable scenes:

1. permutation of any pair of axis,

2. inversion of any of the axis,

3. interchange between any of the Cartesian components of s and r.

These considerations allow the restriction of the space $\mathbf{S}$, parametrized by $\{R, \mathbf{l}, \mathbf{s}, \mathbf{r}\}$, to the subspace $\mathbf{S}_{acoust}$ of acoustically distinct scenes, which can be obtained from S by imposing the following constraints:

1. $l_z \leq l_y \leq l_x$,

2. $0 \leq s_i \leq l_i/2$,

3. $s_i \leq r_i \leq l_i - s_i$.

Summarizing, the SIFA algorithm described here is to infer a scene from $\mathbf{S}_{phys}$, which is, to select 10 coefficientsa $\{R, \mathbf{l}, \mathbf{s}, \mathbf{r}\}_{inferred}$ subject to the aforementioned constraints, such that the acoustics of such scene is as close as possible to that characterized by $\mathbf{IR}_{ref}$.

2.2.2 Heuristic search

This formalism provides a limited search space where an heuristic search can be run given a proper cost function. We have presented, in a conference paper[4], a genetic algorithm that solves that problem, for a limited set of cases as a first step towards more general SIFA algorithms.

The cost of a given solution was computed by simulating the IR and comparing it to the target one. The IR is simulated using the Image Source method [5], which is quite straight forward in rectangular rooms. Both IR's are compared by evaluating following formula:

$$E = 1 - \int_0^{t_f} |IR^L_{cand}(t)||IR^L_{ref}(t)|w(t)dt \tag{1}$$

Were IR^L means the normalized and low pass filtered IR's and $w(t)$ is a positive weighting function. The low pass filter is obtained by convolving the signal by a Gaussian function.

The results of such algorithm are already useful for practical usage, but its convergence, target extension and the goodness of the solutions have been slightly tested. So further testing, facing, for instance, different real scenarios, is required.

3 Extension request

As described in the report there are several areas that require further work.

First, there is a paper we are about to submit about the destructive effect of the Ambisonics decodings. We would like to introduce the eventual

input of the reviewers into the thesis to improve the quality of the thesis. Regarding the results highlighted in that paper, we want to perform measurements the predicted filter effect in real set scenarios. Also, as described in previously, analytical results suggest that several equalization strategies could be applied to compensate the effects of the head. We want to test such equalization strategies and evaluate with user tests.

Regarding SIFA problem, current approach has been tested in a quite weak way. A more exhaustive evaluation is needed to back the results. We need to take measurements of real rooms and test psychoacoustic tests either using objective measurements or user tests. We would also write and submit a paper covering such evaluation.

Finally, the thesis itself must be written. The two topics that have received most of the focus are quite far components of the wider work-flow of relating the audio to the 3D scene, so they must be properly contextualized and connected.

So, considering this scenario, we ask for an extension of six months to the submission date, in order to conclude the research, writing and submitting pending publications, incorporate reviewers feedback into our work and writing the thesis.

References

[1] B. Gardner and K. Martin, "HRTF Meassurements of a KEMAR Dummy-head microphone," MIT Media Lab Perceptual Computing, Tech. Rep. 280, 1994.

[2] "Listen HRTF Database." [Online]. Available: http://recherche.ircam.fr/equipes/salles/listen/

[3] W. Kreuzer and Z. Chen, "A fast multipole boundary element method for calculating hrtfs," in *AES Convention*, 2007.

[4] David García-Garzón and Daniel Arteaga and John Usher and Toni Maneos, "Determining a scene geometry from its impulse response," in *Proceedings of the 2010 Internoise Conference*, 2010.

[5] H. Kuttruff, *Room acoustics*, Applied Science, Ed., 1973.

Barcelona, 23/05/2011

The Ph.D Candidate,
David García Garzón

The Ph.D. Thesis Director,
Prof. Vicente López

The Ph.D. Thesis co-supervisor,
Dr. Toni Mateos